Relevance of Massively Distributed Explorations of the Internet Topology: Simulation Results

Jean-Loup Guillaume, Matthieu Latapy

Relevance of Massively Distributed Explorations of the Internet Topology: Simulation Results

Jean-Loup Guillaume, Matthieu Latapy
LIAFA – CNRS – Université Paris 7
2 place Jussieu, 75005 Paris, France.
(guillaume,latapy)@liafa.jussieu.fr

Abstract— **Internet maps are generally constructed using the traceroute tool from a few sources to many destinations. It appeared recently that this exploration process gives a partial and biased view of the real topology, which leads to the idea of increasing the number of sources to improve the quality of the maps. In this paper, we present a set of experiments we have conduced to evaluate the relevance of this approach. It appears that the statistical properties of the underlying network have a strong influence on the quality of the obtained maps, which can be improved using massively distributed explorations. Conversely, we show that the exploration process induces some properties on the maps. We validate our analysis using real-world data and experiments and we discuss its implications.**

Index Terms— **Network measurements, Graph theory, Simulations.**

INTRODUCTION.

Due to its fully distributed construction and administration, mapping the Internet (in terms of machines and physical links between them) is a challenging task. It is however essential to obtain some information on its global shape. Indeed, it plays a central role in key problems like network robustness [3], [10], [11], simulation of future protocols and uses [30], and many others.

Exploring the Internet topology is a research problem in itself [18], [21], [28], [40], [42]. Indeed, many difficulties (like the identification of the multiple interfaces of a same router) arise when one wants to map the Internet. Various techniques and methods have been introduced to achieve this goal. Some of them are very subtle, but current explorations still rely on the extensive use of the `traceroute` tool: one collects routes from a given set of sources to a given set of destinations, and then merges the obtained paths. Some post-processing is generally necessary to clean the obtained data, but we do not enter in these details here.

Two points are particularly important in the scheme sketched above. First, it must be clear that the image we obtain from the network is *partial* (except if the number of sources and destinations is huge, we certainly miss some nodes and some links) and may be *biased* by the exploration process (some properties of the obtained map may be induced by the way we explore the network). Second, the number of sources cannot be increased easily, whereas one can take as many destination as one wants. Indeed, one needs direct access to the sources in order to run the `traceroute` tool, whereas one only needs the IP addresses of the destinations. In the case of [18], which is one of the largest explorations currently available, only a few dozens of sources are used whereas there is several hundreds of thousands destinations.

Recently, several researchers conduced experimental and formal studies to evaluate the accuracy of the obtained maps of the Internet [9], [23], [24], [25], [36], [39]. All these studies use simple models of networks and `traceroute` but they all give good arguments of the fact that the currently available maps of the Internet are very incomplete, and that there probably is an important bias induced by the exploration process.

In order to improve these maps, several researchers and groups now propose to deploy massively distributed measurement tools [17], [37], [38]. The basic idea is that dramatically increasing the number of sources would significantly improve the quality of the obtained maps. Our central aim in this paper is to rigorously evaluate the relevance of this approach.

To achieve this, we conduce an extensive set of experiments designed as follows. We consider a graph G representing the network to explore. We then simulate the exploration process and obtain this way a (partial and biased) view G' of the original graph. We then compare G' and G to evaluate the quality of this view. We process this simulation using all the possible numbers of sources and destinations, which makes it possible to study the impact of these numbers on the accuracy of the obtained view.

This method is not new: since its introduction in the leading paper [25], it has been used in [9], [23], [12]. However, whereas in these papers the authors consider only one or few sources and study the bias induced on the degree distribution, we will here use a wide variety of numbers of sources and consider a rich set of statistical properties.

This paper is organized as follows. First we define the statistical properties of networks relevant to our study, we present the models we use and discuss our methodology (Section I). Then we present and analyze the results of our simulations on various models and statistical properties (Sections II–III). Section V is devoted to the comparison of our results with real-world data and experiments, which makes it possible to identify the most meaningful simulations. Finally we present our conclusions and discuss them.

I. PRELIMINARIES.

A network topology can naturally be represented by a graph. For our purpose, the graph does not need to be weighted nor directed. A route in the network, as given by the `traceroute` tool, is a path in the corresponding graph. Since a few years [8], [16], [18], [34], a strong effort has been made to discover the topology of the Internet by extensive use of `traceroute` and other tools (BGP tables, source routing, etc).

The obtained maps give much information on the global architecture of the Internet. In particular, they gave evidence of the fact that the Internet topology has some statistical properties which make it very different from the models used until then [7], [16]. This induced an intense activity in the acquisition of such maps [18], [21], [34], in their analysis [16], [41] and in the accurate modeling of the Internet [6], [29], [43], [44]. See [35] for a survey.

Our analysis of the exploration process will be based on these statistical properties and these models, which we present below. We also need to model the `traceroute` tool and the exploration process, which we also discuss in this section. Finally, we present our methodology, and explain how our results should be read.

Statistical properties

The Internet, at router level, is composed of several millions of nodes and dozens of millions of links. Let N denote its number of nodes and M its number of links.

It is well known and quite intuitive, that the density of the Internet graph is low: the number of existing edges over the number of possible ones, $\frac{2 \cdot M}{N \cdot (N-1)}$, is low. In other words, the average degree k of the nodes (their average number of links), *i.e.* $k = \frac{2 \cdot M}{N}$, is a constant independent of the size of the network.

A less known point is that the average distance (length of a shortest path between two nodes) is low. It typically scales as $\log(N)$. This is however not surprising, since it is an essential objective of the design of the network, and since it is actually very natural for any graph to have a low average distance [5], [26], [33].

On the contrary, although it is now well understood, the fact that the degree distribution of the Internet graph follows a power law has been a surprise [16]. Indeed, the proportion p_k of nodes of degree k scales as a power of k: $p_k \sim k^{-\alpha}$ with $\alpha \simeq 2.5$. Intuitively, this means that most nodes have a low degree but there exists some nodes with (very) high degree. Such graphs are said to be *scale-free*.

Another important statistical property measured on the Internet is its clustering C defined as $C = \frac{N_\Delta}{N_\vee}$, where N_Δ is the number of triangles (three nodes with three links) in the network and $N_\vee$ is the number of connected triples (three nodes with at least two links). In other words, C is the probability that two nodes are connected together, given that they are both connected to a same third, which gives a measure of the local density of the graph. The clustering of the Internet is high, considered as a constant independent of N.

All these claims (low density, low average distance, power law degree distribution and high clustering) follow the opinions most widely spread today. However they rely on measurements processed on partial and biased views of the actual Internet. They should therefore be considered carefully. In particular there is a lot of discussion about the presence of power-law degree distribution [7].

Modeling networks

The basic model for networks is the Erdos and Rényi (ER) random graph model [5], [15]. In an ER graph with n nodes, each of the $\frac{n \cdot (n-1)}{2}$ possible links exists with a given probability p. In other words, an ER graph is constructed from n nodes by choosing $m = p \cdot \frac{n \cdot (n-1)}{2}$ links at random. Notice that an ER graph contains a giant component as soon as the average degree is greater than 1. In the following this condition is always fulfilled an generally the graph itself is fully connected.

In such a graph, the average distance grows as $\log(n)$ [5] as long as p is high enough. However, the clustering is small (it tends to zero when n grows), and the degree distribution follows a Poisson law ($p_k \sim e^{-\alpha} \frac{\alpha^k}{k!}$). This implies in particular that all the nodes have a degree close to the average. Therefore, although this model can be considered as relevant concerning the average distance, it misses the two other main properties of the Internet.

An important step was made when Albert and Barabási (AB) introduced their model based on *preferential attachment* [1], [14]. In this model, nodes arrive one by one and choose k neighbors among the existing ones with a probability proportional to their degree. The degree distribution of the nodes in the obtained graphs follow a power-law with an exponent -3 (it is possible to modify this exponent in others models using preferential attachment). The average distance of such a graph is logarithmic in the number of nodes, but the clustering is low.

This model has been modified to give highly clusterized graphs: in the Dorogovtsev and Mendes (DM) model [13], nodes arrive one by one but at each step one chooses a random *link* $\{u, v\}$ and the new node is linked to both u and v. This implies that a node is chosen with a probability proportional to its degree. Therefore, the preferential attachment principle is hidden in this model, which induces the fact that DM graphs have a power-law degree distribution. Moreover, since one forms a triangle at each step, they have a high clustering.

It is also possible to sample a random graph with a prescribed degree distribution using the Molloy and Reed (MR) model [27], [31], [32]. This gives graphs with exactly the wanted degree distribution, but with low clustering.

Finally, the Guillaume and Latapy (GL) model [22], based on bipartite graphs, gives graphs with power law degree distributions and high clustering, by sampling graphs with prescribed distribution of clique (complete sub-graph) sizes.

These models are currently the most widely used for the realistic modeling of clusterized scale-free networks and have all their own advantages. In particular, the parameters are different from one model to another: the main parameter for ER and AB models is the average degree, and the others properties of these models (the degree distribution for instance) are consequences of the construction process itself. Likewise, the original DM model has no parameter but the size of the generated graph

and once again, the properties of this model are contained in the construction process. Finally, MR and GL models are defined using the degree distributions one wants to obtain, and most of the properties (including the average degree) are consequences of these distributions. Therefore, depending on the objective (degree distribution, clustering, etc), one will use one model rather than another. These models have been considered as building blocks for more complex models. See [2] for a description of some of these.

In the results we present here, our aim is to give evidence of the impact of network properties on the efficiency of shortest-path based explorations. In most cases, the results do not vary qualitatively between the AB and the MR model on the one hand (which have a power-law degree distribution and no clustering), and between the DM and the GL ones on the other hand (both power-law degree distribution and clustering). We will therefore mainly present results on ER, AB and DM models, except in Section V where it is particularly relevant to use MR and GL ones.

Modeling `traceroute` *and the exploration*

In this paper, we will make the classical assumption [12], [25], [23] that a route as obtained by `traceroute` is nothing but a shortest path between the source and the destination. It is known that this is not always true [19], [24], but the realistic modeling of routes is nowadays an open problem.

Moreover, let us emphasize on the fact that we will make an intensive use of routes simulations, which makes it crucial to be able to process them very efficiently. To this respect, our assumption has important advantages.

Since there may be many shortest paths between two nodes, this is not sufficient to properly define a model of `traceroute`. At a given moment, the route followed by a packet when a given router R routes it to a destination D will always be the same independently of the sender. This may have an influence on the quality of the exploration process, therefore we included it in our model of `traceroute`: we always follow the same shortest path (initially chosen randomly) between any two nodes. In [25] a similar definition of `traceroute` based on shortest-paths has been introduced.

We now have a precise model of routes as viewed by `traceroute`. But we also need a model for the exploration process. We considered two points of view: in the first one we suppose we make a *snapshot* of the network, and in the second one we suppose we make a *long-time* exploration. This leads respectively to the *unique shortest path* (USP) model, and to the *all shortest paths* (ASP) one: we either see only one route for any given source and destination, or we see all the possible ones. The ASP model should not be considered as a realistic model, since one cannot expect to get all shortest-paths even within a long period of time (in such a long time, the network is very likely to evolve). However it can be considered as a *best case* when dealing with shortest-paths or as an upper bound on the amount of information on can expect from a shortest-paths based exploration.

We also conduced experiments using other models (random shortest path, several shortest path but not all, etc), but the results do not qualitatively vary, so we do not detail them here.

Finally, we generally consider a set of sources and a set of destinations, and make the exploration using each possible couple of source and destination in these sets. Such a model has already been used in [4], [25], where the authors call it a (k, m)-`traceroute` study (k is the number of sources and m the number of destinations).

Methodology

Following [25], our global approach is as follows:

1) generate a graph G using a given model with some known parameters,
2) compute a view G' of G using a given model of the exploration process and a set of sources and of destinations, and
3) compare the statistical properties of G' to the ones of G.

Let us insist on the fact that we seek *qualitative* results only: we want to know how qualitative properties of the network influences the properties we observe during an exploration process, and how reliable are the obtained maps with respect to some statistical properties. It makes no sense to interpret quantitatively the results obtained with the kind of approach we use here. On the contrary, by the simplicity of the models and of the properties we use, we obtain evidences of the fact that some properties play a fundamental role in the exploration whereas others may be neglected.

In the method sketched above, the third point (comparison of the original graph with the view we obtain) is a difficult task. To achieve it, we will make an extensive use of grayscale plots defined as follows (see Figure 6 for some easily readable examples). For a graph G of N nodes, we consider a square of size $N \times N$. Each point (x, y) of the square corresponds to a view G' of G using x sources and y destinations with a given model of the exploration process. The point is drawn using a grayscale representing the value of the real-valued statistical property p under consideration: from black for $p = 0$ to white for the maximal value obtained for p (which might be greater than its value for G).

Therefore, in these plots, the point $(0, 0)$ is always black (we do not see anything using zero sources and zero destinations) and the point (N, N) has the grayscale corresponding to the value of p for the original graph G (when every node is a source and a destination, we see everything: $G' = G$). The points darker than the point (N, N) correspond to conditions where the value of p is under-estimated, whereas points clearer correspond to conditions where it is over-estimated. The white points correspond to the maximal values reached by p. Notice also that the gray variation is linear: if a dot is twice darker than another dot, then the associated value is twice as large.

Each point of such a plot corresponds to a graph G', and therefore computing such plots is computationally expensive. Therefore, is it important to efficiently compute them and to keep N quite low. We conduced experiments with $N = 10^3$, $N = 10^4$ and $N = 10^5$ typically, and, whereas some finite size effects are visible on small graphs ($N = 10^3$), these effects disappear for graphs of size $N = 10^4$ and more. This is why we will present plots for this value of N in general.

Finally, to improve the grayscale plots readability, we added on each such plot the 0.25–, the 0.50–, the 0.75– and the 0.99–

level lines, where the l–level line is defined as the set of points where the value of p over its maximal value is between $l - 0.01$ and $l + 0.01$. These lines are often a precious help in the interpretation of the grayscale plots. See Figure 6 and the rest of the paper for examples.

II. Proportion discovered

In this section, we focus on the most basic statistical properties of an exploration, namely the proportion of discovered nodes, the proportion of discovered links, and the quality of the evaluation of the average degree. We present the relevant results on the ER, the AB, the MR and the DM models, and we explain which parameters have a strong influence on these results.

Notice that results using similar approach have been obtained in [4], however our explorations are processed on random graphs instead of real data, the aim being to highlight the parameters of the models and therefore the characteristics of the graphs which influence the efficiency of the exploration.

Random graphs

Let us first study what happens during the exploration of an ER graph. Figures 1 and 2 plot the proportion of the graph discovered. When the average degree is quite small, there is no qualitative difference between ASP and USP (there exists in general very few shortest path between any two nodes) and the quality of the view is good even for small numbers of sources and destinations (Figure 1 shows the USP plots, which are very similar to the ASP ones in this case).

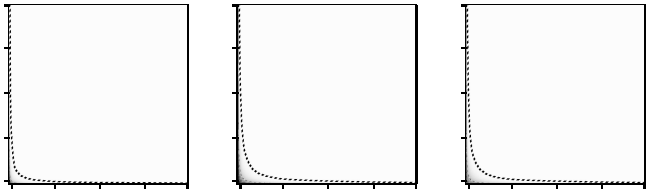

Fig. 1. ER graph: number of vertices, number of edges, and average degree. $k = 10$, $N = 10^4$, USP. The ASP plots are very similar in this case.

On the contrary, when the average degree grows, so does the number of shortest paths, and the difference between ASP and USP becomes significant. This can be observed in Figure 2, where we show the plots for both USP and ASP on an ER graph with high average degree. In this case, the vertices are not harder to find than in a low-average degree graph, but the edges are.

The fact that the average degree is obtained by dividing two other properties which are improved by the use of more sources and/or destinations has important consequences. If one of the two properties is highly biased and the other is not, then the average degree will have a strong bias. The quotient acts like a *worst case* filter. Figure 2 shows this effect on dense ER graphs. Since the number of edges is very poorly estimated, so is the average degree.

Notice however that when N grows, the proportion of sources and destination necessary to obtain an accurate view decreases, even if the number of sources and destinations increases.

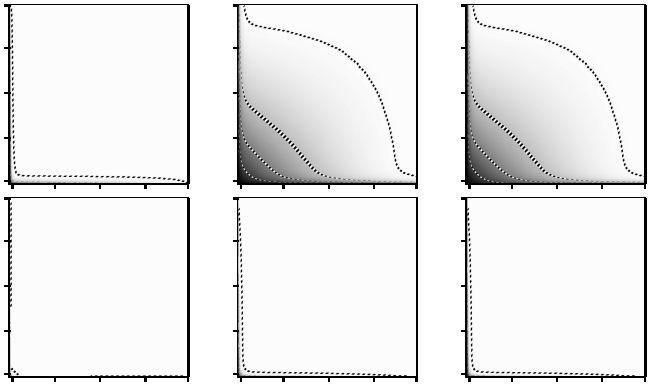

Fig. 2. ER graph: number of vertices, number of edges, and average degree. $k = 100$, $N = 10^4$, USP (first line) and ASP (second line).

Scale-free graphs

Let us now observe what happens when we consider scale-free graphs. Let us begin with the AB model which makes it possible to obtain scale-free graphs with a given average degree (by choosing the number of edges created for each new vertex). In Figure 3 (all the plots, using different parameters, display a very similar behavior), we can see that the efficiency of the exploration on such graphs is qualitatively similar to the one on ER graphs, though it is lower. If we want a very precise map, however, we need much more sources and destinations. There is also a strong difference between USP and ASP, which tends to show that there are multiple shortest paths between nodes.

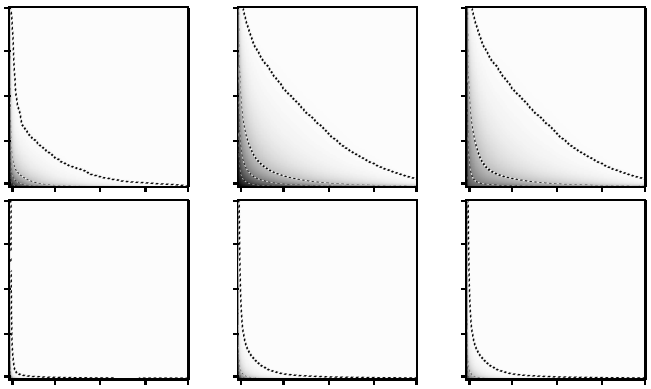

Fig. 3. AB graph: number of vertices, number of edges, and average degree. $k = 10$, $N = 10^4$, USP (first line) and ASP (second line).

If we make the same experiments with MR graphs, which also have a scale-free nature and should be equivalent to AB graphs, we obtain the surprising results plotted in Figure 4: the quality of the obtained view is much worse for MR graphs than for AB graphs. Even when considering ASP, one needs to take about half sources and destinations to view 75% of the graph (both in terms of edges and nodes).

Notice also that the average degree is surprisingly well estimated, even if overestimated. Since the average degree is the quotient of the proportion of nodes and edges discovered, if the two properties has the same kind of bias, this may be hidden by the quotient: the evaluation of the average degree is good whenever the ratio between the number of edges and the number of

nodes is accurate, even if these numbers themselves are wrong. Figure 4 plots such a behavior. Actually the average degree is overestimated since high degree nodes and some of the edges attached to them are first discovered and low degree nodes are discovered only in the later steps of the exploration.

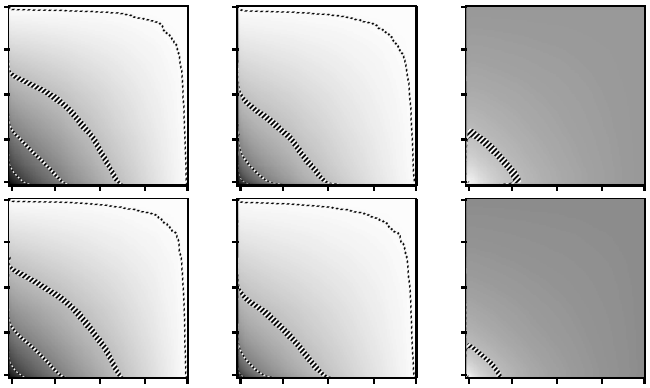

Fig. 4. MR graph: number of vertices, number of edges, and average degree. $\alpha = 2.5$, $N = 10^4$, USP (first line) and ASP (second line).

The fact that MR graphs are harder to explore than AB ones rely on a simple explanation of this fact: in an AB graph with average degree k, the minimal degree is $\frac{k}{2}$ (we add $\frac{k}{2}$ links at each step, see Section I). Therefore, the power-law degree distribution of such a graph stands only for nodes with degree higher than $\frac{k}{2}$. On the contrary, in a MR graph, the number of low-degree nodes (and in particular the number of nodes with only one link) is very high. During an exploration process, these nodes are difficult to discover since they lie on very few shortest paths. For example, a node of degree 1 and the link attached to it are discovered only when we choose this node as a source or destination. If the number of such nodes is high then the estimation of the size of the graph can only be good with a lot of shortest paths.

These explanations can be checked as follows. Instead of considering the original MR graph, we consider its *core* defined as the graph obtained by removing all the nodes of degree 1 and iterating this process until there is no such node anymore. In other words, the graph is composed of the core, to which are attached some tree-like structures, which we remove. If we run the exploration on the core of a MR graph, we obtain the plots in Figure 5. For the USP exploration, these results are more in accordance with the ones for the AB graphs. Notice however that it is not only difficult to find a node of degree 1, but also to find all the nodes of low degree, which explains the difference between AB (no nodes of degree lower than $\frac{k}{2}$) and the core MR graphs.

The difference between ASP and USP is more important in AB graphs than in MR (or in the core of MR), which shows that there are more multiple shortest paths in an AB graph than in a MR one.

The important point here is that the quality of an exploration of a MR graph is low because of the large number of low-degree nodes. Such nodes, among which are tree-like structures, are difficult to discover since they lie on few shortest paths, whereas the core of the graph and especially the nodes of high degree are rapidly discovered.

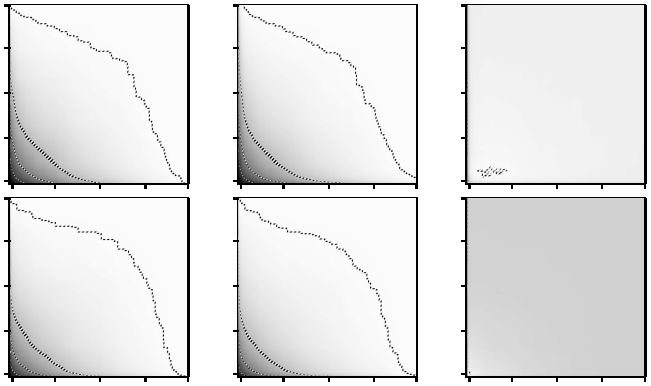

Fig. 5. Core of a MR graph: number of vertices, number of edges, and average degree. $\alpha = 2.5$, $N = 10^4$, USP (first line) and ASP (second line).

Clusterized graphs

Let us now consider a DM graph, in which there are many triangles and the degree distribution follows a power law. Like in an AB graph, there is no node with only one link. Therefore, the effect noticed above in MR graphs should not appear.

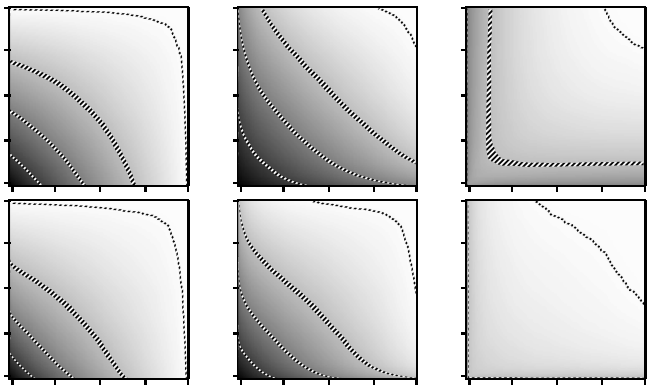

Fig. 6. DM graph: number of vertices, number of edges, and average degree. $N = 10^4$, USP (first line) and ASP (second line).

However, one can see in Figure 6 that we again obtain low quality maps of this kind of graphs. The fact that the plots for USP and ASP are very similar indicates that there are very few different shortest paths between nodes. This, and the fact that the quality of the obtained views is low, can be understood as follows. When one wants to explore a clique (complete graph), or more generally a dense graph, one has to use a large number of sources and destinations. For instance in a simple triangle, two edges cannot be discovered simultaneously by one `traceroute`. Therefore three `traceroute` have to be processed to discovered a triangle. The same happens for a k-clique in which $k \cdot (k-1)/2$ `traceroute` have to be processed. The high clustering in DM graphs is equivalent to the fact that there are many subgraphs which are cliques or almost. All these parts of the graph are difficult to explore.

Notice that this time the average degree is poorly estimated, which shows that inferring the average degree is very sensitive: very similar behaviors (see Figures 4 and 6 for instance) may

lead to very different average degree estimations. This warns us against drawing precise conclusions for the average degree from such explorations.

Finally, the conclusion of this section is the following: two properties of graphs make them hard to explore in different ways. The first one is the large number of tree-like structures around the core of the graph. The second one is the high clustering which induces many dense subgraphs. The two properties are complementary and act on different parts of the graph (on the border and on the core, respectively), which indicates that we should take them both into account.

III. Degree distribution

The degree distribution of the Internet has recently received much attention. It is the main property for which the bias induced by the exploration have been studied [9], [23], [24], [25], [36], [39]. In particular in [25] it is shown that under simple assumptions it is possible to obtain a view with an heavy tailed distribution from an ER graph. We will deepen these study here by considering several models, exploration methods, and numbers of sources and destinations. However, we cannot use the grayscale plots in this context, since the question we address is: how fast does the observed degree distribution converge to the real one with respect to the number of sources and destinations? This can not be directly evaluated as a real number which would be necessary for grayscale plots. Instead, we display plots for representative values of the parameters (again, we conduced extensive simulations but we selected the most relevant ones for this presentation).

Random graphs

Let us first consider ER graphs with low average degree. As shown in Figure 7, if the number of sources is very low then the obtained degree distribution is far from the real one. With an USP exploration, the obtained degree distribution converges quite slowly: it is still significantly different from the real one if we take 1% of sources and 10% of destinations. With an ASP exploration, the accuracy is much better: the view is almost perfect even with only 0.5% of sources and 20% of destinations.

The case of ER graphs with high average degree (Figure 8) is more interesting: the presence of high degree nodes makes it possible to obtain power-law degree distributions with partial USP explorations. This has been studied in previous works [25], [36] to show that the exploration bias may be qualitatively significant. This measurement bias occurs when one considers very few sources and many destinations (Figure 8, top) and the USP exploration. It disappears when one considers a larger number of sources, for instance 0.5% of the whole (Figure 8, bottom), or when one considers an ASP exploration (Figure 9), even for small numbers of sources and destinations.

Notice also that, in intermediary cases, one may obtain surprising results like the plot for 500 sources and 5000 destinations in Figure 8, which has two peaks. As explained in [25], this is due to the fact that in such cases most of the links close to the sources are discovered, whereas the ones close from the destination are not. The rightmost peak then corresponds to nodes close from the sources (for which we have all their edges) while the leftmost one corresponds to the nodes close from the destinations (for which we miss almost every link).

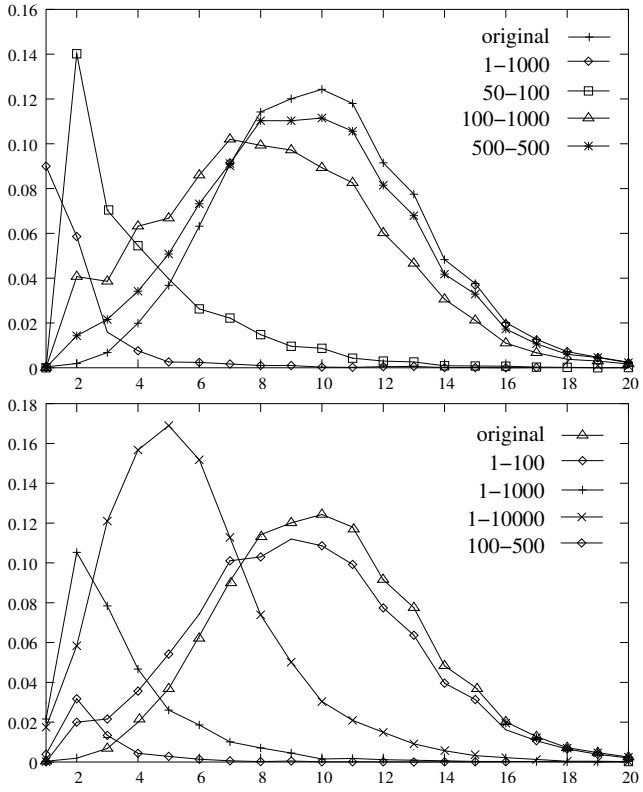

Fig. 7. ER graph: degree distribution. $k = 10$, $N = 10^4$, USP (top) and ASP (bottom).

These first results concern ER graphs, for which the degree distribution are not power-laws. They show that it is quite difficult to obtain an accurate view of the degree distribution of such graphs, which is improved significantly by the use of many sources and destinations. As already noticed, the use of a low number of sources may even give degree distributions qualitatively different from the real ones.

Scale-free graphs

If we now consider scale-free graphs, the results are totally different: as one can check in Figures 10 and 11 respectively for MR and DM graphs, USP explorations give accurate views of the actual degree distribution [1], even for small numbers of sources and destinations. In the case of MR graphs (the results are the same for AB graphs), the fit is excellent. In the case of DM graphs, the obtained exponent is slightly lower for small numbers of sources but it rapidly converges to the real one.

In conclusion, the behaviors of ER and scale-free graphs are completely different concerning the accuracy of the obtained degree distributions. Whereas it is quite difficult (especially using an USP exploration) to obtain an accurate estimation for

[1] The important characteristic of a power-law distribution is its exponent α, *i.e.* the slope of the log-log plot. Here, to improve the plots readability, we divide the number of nodes of a given degree by the total number of nodes N, including the ones which are not discovered during the exploration in concern. This does not change the slope α

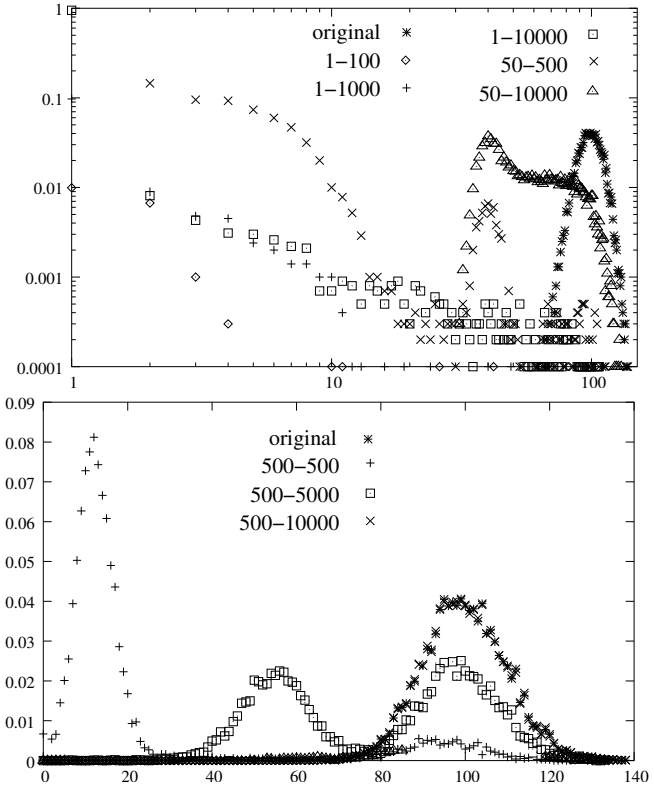

Fig. 8. ER graph: degree distribution. $k = 100$, $N = 10^4$, small number of sources (top log-log scale) and large number of sources (bottom normal scale). USP

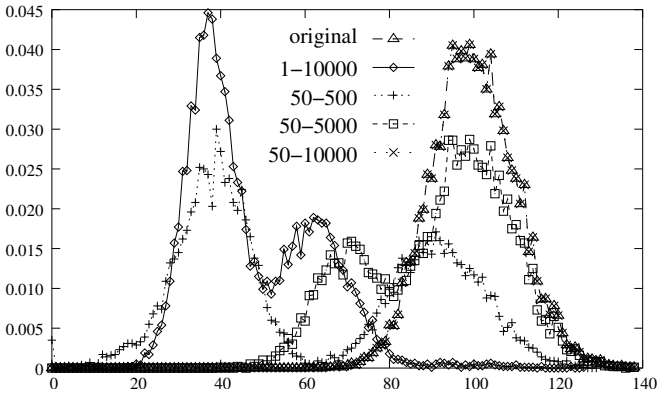

Fig. 9. ER graph: degree distribution. $k = 100$, $N = 10^4$, ASP.

ER graphs, the exponent of the power-law degree distribution of a MR, an AB or a DM graph is correctly measured even with a small number of sources and destinations. We also show that, despite the fact that using a very small number of sources and a large number of destinations can in principle give us a wrong idea of the actual degree distribution of a graph, these cases are quite pathological.

IV. Clustering

The clustering of a graph is computed by dividing the number of triangles in the graph by the number of connected triples (see Section I). Just like the average degree depends on the obtained numbers of nodes and links (see Section II), this means

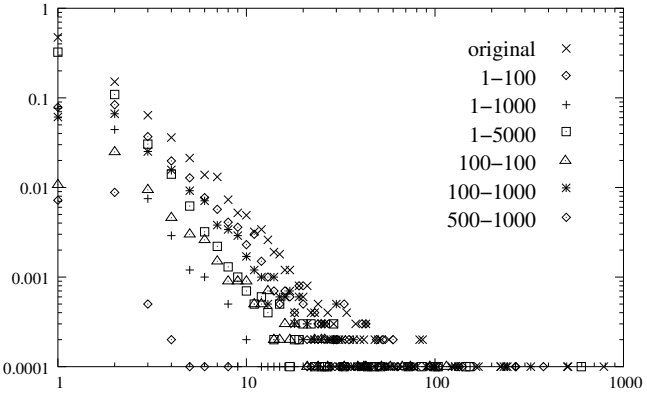

Fig. 10. MR graph: degree distribution. $\alpha = 2.5$, $N = 10^4$, USP.

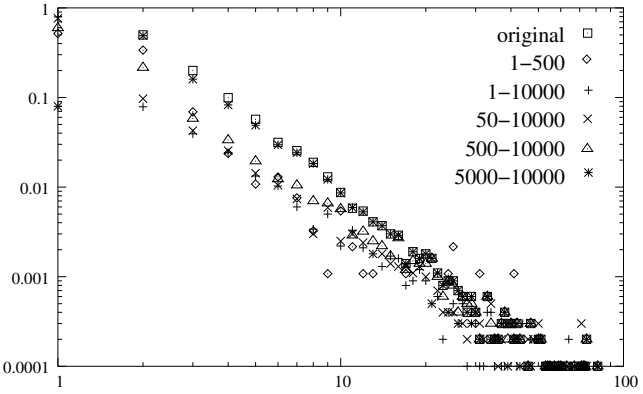

Fig. 11. DM graph: degree distribution. $N = 10^4$, USP.

that the evaluation of the clustering of a graph we obtain using an exploration depends on how fast we discover triangles with respect to the speed at which we discover triples: the evaluation of the clustering is accurate if we discover a proportion of the total number of triangles similar to the proportion of the total number of triples we discover. We will therefore study how triangles and triples are discovered, together with the clustering itself.

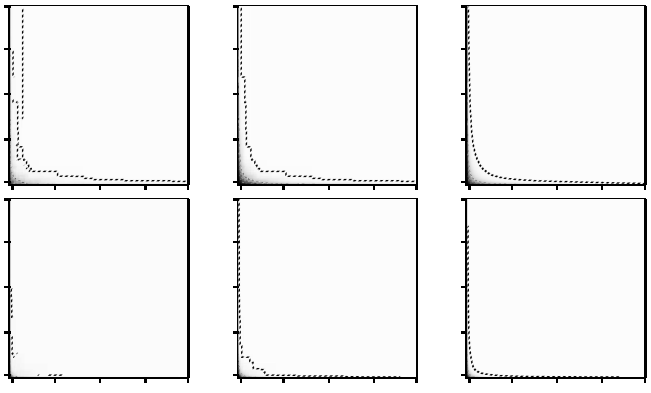

Fig. 12. ER graph: clustering, number of triangles, and number of triples. $k = 10$, $N = 10^4$, USP (first line) and ASP (second line).

Let us first observe what happens for ER graphs. Notice that when the average degree is low, there are almost no triangles in such graphs (and so the clustering is zero). When the average

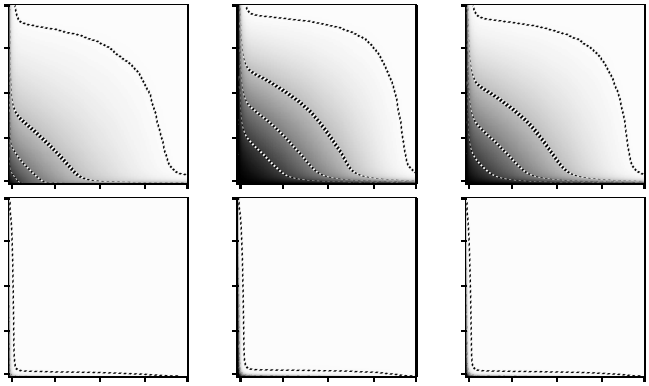

Fig. 13. Dense ER graph: clustering, number of triangles, and number of triples. $k = 100$, $N = 10^4$, USP (first line) and ASP (second line).

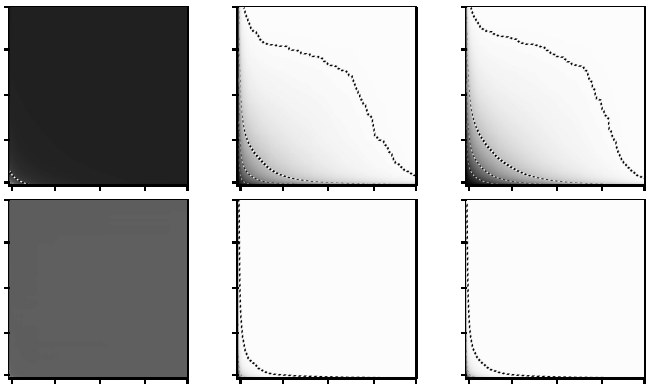

Fig. 14. AB graph: clustering, number of triangles, and number of triples. $k = 10$, $N = 10^4$, USP (first line) and ASP (second line).

If we turn to AB and MR graphs (the behaviors of the two kinds of graphs are very similar), we again have a very low clustering but in the USP case it is over-estimated when we consider few sources and destinations. This is a consequence of the fact that we discover much more triangles than triples at the very beginning of the exploration. However, the estimations rapidly becomes accurate, and lower than the initial value. This can be seen in Figure 14: the black value corresponds to the clustering of the original AB graph, and the only cases where the estimation is wrong are in the lower left corner. The ASP explorations give more accurate results.

Let us now observe what happens with a highly clusterized graph, obtained with the DM model. In Figure 15, we can see that the clustering can be well evaluated if we use as many sources as destinations. If we use much more sources than destinations or conversely then the estimation is bad (notice that this is currently the case for the explorations of the Internet).

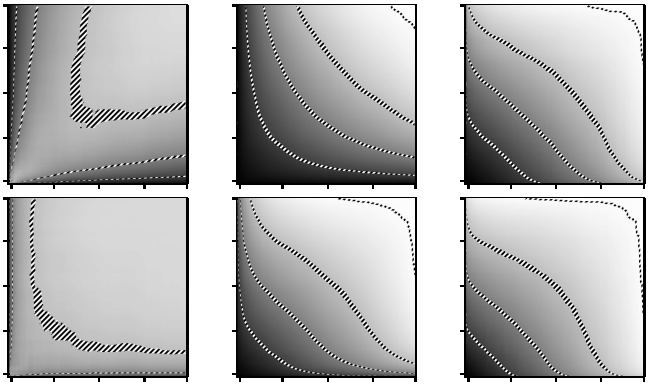

Fig. 15. DM graph: clustering, number of triangles, and number of triples. $N = 10^4$, USP (first line) and ASP (second line).

Indeed, in these cases, there is a strong difference between the rapidity with which we discover triangles and triples. When the numbers of sources and destinations are similar, on the contrary, despite we miss many triangles and triples, the proportions we miss of each are similar. In this case, therefore, the estimation of the clustering is accurate.

In conclusion, we see in this section that the clustering may have a lot a different behaviors since it is computed by a ratio of two parameters: triangles and triples (it is quite similar to the average degree), and even if both parameters are not well estimated, the clustering itself may be. Discovering dense subgraphs and in particular triangles is never an easy task using shortest-paths, however discovering triples might not be very easy in dense graphs since they often belong to triangles (see Figure 13 and 15), which means that two links of the triple cannot be discovered with only one path. In all cases, increasing the number of sources and destinations gives a better approximation of the clustering.

V. Real-world data and experiments

Until now, we presented simulations carried out on models of networks and using simple models for `traceroute` and the exploration process. We will now make the same kind of experiments on real-world data to evaluate the relevance of these simulations.

To achieve this, we use the *core* of the *Mercator* map of the Internet [20], [21], *i.e.* the subgraph obtained by iteratively removing the nodes of degree 1 from the original *Mercator* map. This map has all the properties we have mentioned: low density, high clustering, power-law degree distribution and low average distance. Notice that this map has been obtained with only one source and use of source-routing and therefore the protocol cannot be compared with the one used in our study. However we only study the map itself and consider it representative of the Internet.

We restrict our study to the core of this map because we have already seen that the tree-like structures around it are difficult to discover, and our aim is now to identify other properties which may influence the exploration. Using this graph, we make exactly the same measurements as the ones presented above and we compare the results with the ones obtained on a random

graph having exactly the same degree distribution (MR model) and on graphs having the same distribution of clique sizes (GL model). See Figure 16 for the basic statistics, and Figure 17 for the clustering [2]. The results concerning the average distance and the degree distributions are similar to the ones observed on models, therefore we do not discuss them further.

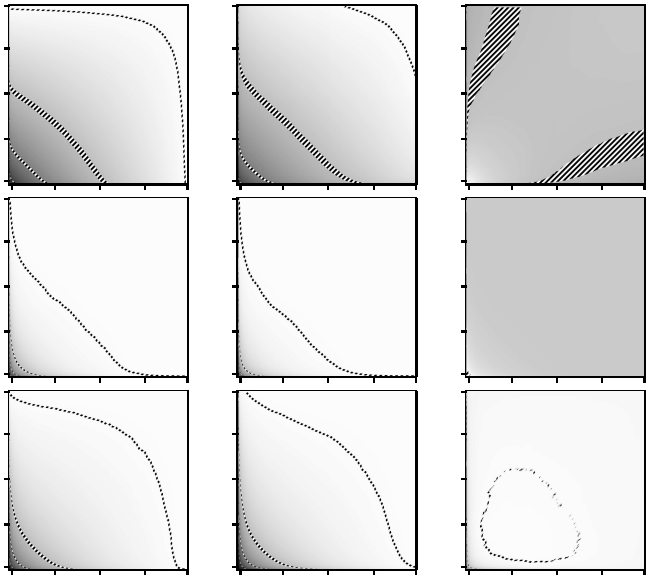

Fig. 16. Number of nodes, number of links, and average degree for (from top to bottom): the original *Mercator* graph, a MR graph with exactly the same degree distribution, and GL graph with the same distribution of cliques sizes. USP explorations.

From Figure 16 and the ones discussed before, we can derive the following observations:

- the low quality of the exploration of the *Mercator* graph is not only due to the presence of tree-like structures around the core, since we removed them in this experiment,
- the *Mercator* graph cannot be viewed as a MR graph since the exploration of its core gives results different both from the explorations of the core of a MR graph (Figure 5) and from the explorations of MR graphs with the same degree distribution (Figure 16, second line),
- the clustering could be viewed as the main property responsible for the low quality of the explorations, since the results for the *Mercator* graph are very similar to the ones for DM graphs (Figure 6, first line) and quite similar to the ones for GL graphs (Figure 16, third line).

This last conclusion, however, is not completely satisfactory. Indeed, the results concerning the quality of the estimation of the clustering are significantly different for DM graphs (Figure 15) and for the *Mercator* graph (Figure 17, first line). The clustering certainly plays a role in the exploration of the *Mercator* graph, but it is much more similar to the one observed for GL graphs (Figure 17, second line). It therefore seems that the models do not capture all the properties which influence the exploration process, even if the low degree nodes and the clustering have been clearly identified among them.

[2]The jumps in the grayscale plots for the clustering of the *Mercator* graph are due to the ones in the plot of the number of triples. Themselves are consequences of the fact that, at this point, we take a very high-degree node as a source with many destinations, which suddenly increases the number of triples (by $d(d-1)$ where d is the degree of the node).

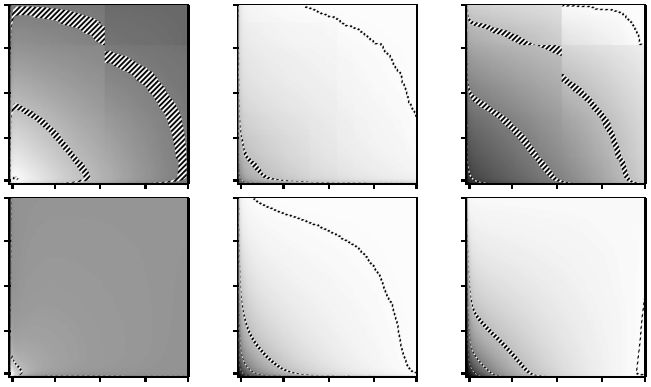

Fig. 17. Clustering, number of triangles, number of triples for the original *Mercator* graph (first line) and a GL graph with the same distribution of cliques sizes (second line). USP explorations.

The exact sources and destinations, and the obtained routes, used to produce the *Mercator* graph are not available. Moreover, it relies on one source and source-routing. Therefore, we cannot plot the grayscale plots where we would take the same sources and destinations as in the *real* exploration, and where we would take real routes rather than shortest paths. Such experiments are currently in progress and we will present them in the full version of this paper.

Conclusion and discussion

We conduced an extensive set of simulations aimed at evaluating the quality of current maps of the Internet and understanding how to distribute explorations massively to improve it. To achieve this, we considered the most commonly used models of graphs (namely the ER, the AB, the MR, the DM and the GL ones). Following the method introduced in [25], we then constructed *views* of these graphs and compared them to the original graphs. We focused on the proportion of the graph discovered (both in terms of nodes and links), the average degree, the average distance, the degree distribution and the clustering, which are the basic statistical properties of complex networks in general, and of the Internet in particular.

We presented in this paper the most significant results. To do so, we introduced the grayscale plots and the level lines, which make it possible to give a synthetic view of a huge amount of information, and to interpret it easily. We also compared the results on network models to the ones obtained on real-world data. This last point confirmed that the simplifications and assumptions we have made in our simulations do not influence significantly the obtained results.

From these experiments, we can derive the following conclusions:

- Two statistical properties of graphs influence strongly our ability to obtain accurate views of them: the presence of many tree-like structure and the high clustering. These two properties act independently and their effects are combined in the case of the Internet.

- It is relevant to use massively distributed exploration schemes to obtain accurate maps of scale-free clusterized networks like the Internet, in particular if we want to discover most nodes and edges, and have an accurate estimation of the clustering. Using more than a few sources should yield much more precise maps.
- On the contrary, the evaluation of the degree distribution of such a network, as well as its average distance (results not presented here) is achieved with very good precision even for reasonably small number of sources and destinations.
- The details of the exploration scheme (for instance USP versus ASP or the behavior of `traceroute`) tends to have little importance when the number of sources and destinations grows. In the case of the Internet, this means that distributing explorations can be viewed as a way to improve the independence of the results from the exploration scheme.
- Despite power-law degree distribution and high clustering play a role in the efficiency of the explorations of the Internet, it seems that other unidentified properties also influence this efficiency.
- Some results not presented here show that it may be relevant not to place sources and destinations randomly in the graph. More surprisingly, the placement of sources and destinations has not the same influence on all the properties.

Finally, these results make it possible to conclude that we may be confident in the fact that the Internet graph has a degree distribution similar to a power-law and that the current evaluation of the exponent of this distribution is quite accurate: current explorations use sufficiently many sources to ensure that we do not obtain biased explorations of ER-like graphs, and in the other cases it seems that the estimation of the degree distribution is accurate. Likewise, one might give credit to the available evaluations of the average distance in the Internet. On the contrary, despite the clustering of the Internet is certainly quite high, the estimations we have should be considered more as qualitative than quantitative.

More investigations are currently in progress. First, we are considering more subtle statistical properties, like the correlations between node degrees, or the correlations between degree and clustering, and more realistic models of `traceroute`. We are also studying real explorations of the Internet using `traceroute` from many sources to many destinations in order to create grayscale plots from real paths.

Finally, let us insist on the fact that most real-world complex networks, like the World Wide Web and Peer to Peer systems, but also social or biological networks are generally not directly known. Various exploration schemes are used to infer maps of these networks, which may influence the vision we obtain. The metrology of complex networks is therefore a general scientific challenge, for which the goal is to be able to deduce properties of the real network from the ones observed.

REFERENCES

[1] R. Albert and A.-L. Barabási. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics 74, 47*, 2002.

[3] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance in complex networks. *Nature*, 406:378–382, 2000.

[4] Paul Barford, Azer Bestavros, John Byers, and Mark Crovella. On the marginal utility of network topology measurements. In *ACM SIGCOMM Internet Measurement Workshop 2001*, San Francisco, CA, November 2001. ACM SIGCOMM.

[5] B. Bollobás. *Random Graphs*. Academic Press, 1985.

[6] T. Bu and D. Towsley. On distinguishing between internet power law topology generators. In *INFOCOM*, 2002.

[7] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. The origin of power laws in internet topologies revisited. In *INFOCOM*, 2002.

[8] K. Claffy, T. Monk, and D. McRobb. Internet tomography. Nature Magazine, Web Matters. http: //helix.nature.com/webmatters/tomog/tomog.html.

[9] A. Clauset and C. Moore. Traceroute sampling makes random graphs appear to have power law degree distributions. cond-mat/0312674.

[10] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Resilience of the internet to random breakdown. *Phys. Rev. Lett.*, 85:4626–4628, 2000.

[11] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Breakdown of the internet under intentional attack. *Phys. Rev. Lett.*, 86:3682–3685, 2001.

[12] L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vasquez, and A. Vespignani. A statistical approach to the traceroute-like exploration of networks: theory and simulations. cond-mat/0406404.

[13] S.N. Dorogovtsev and J.F.F. Mendes. Evolution of networks. *Adv. Phys. 51, 1079-1187*, 2002.

[14] S.N. Dorogovtsev, J.F.F. Mendes, and A. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett. 85*, pages 4633–4636, 2000.

[15] P. Erdös and A. Rényi. On random graphs I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[16] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[17] Cooperative Association for Internet Data Analysis. http://www.caida.org/.

[18] Cooperative Association for Internet Data Analysis Skitter tool. http://www.caida.org/tools/measurement/skitter/.

[19] T. Friedman, M. Latapy, J. Leguay, and K. Salamatian. What is a route on the internet. preprint.

[20] Internet Maps from Mercator. http://www.isi.edu/div7/scan/mercator/maps.html.

[21] R. Govindan and H. Tangmunarunkit. Heuristics for internet map discovery. In *IEEE INFOCOM 2000*, pages 1371–1380, Tel Aviv, Israel, March 2000. IEEE.

[22] J.-L. Guillaume and M. Latapy. Bipartite graphs as models of complex networks. preprint - http://www.liafa.jussieu.fr/ latapy/Publis/, 2004.

[23] J.-L. Guillaume and Matthieu Latapy. Complex network metrology, 2004. preprint - http://www.liafa.jussieu.fr/ latapy/Publis/.

[24] Y. Hyun, A. Broido, and K. Claffy. Traceroute and BGP AS path incongruities. http://www.caida.org/outreach/papers/2003/ASP/.

[25] A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling biases in IP topology measurements. In *IEEE INFOCOM*, 2003.

[26] L. Lu. The diameter of random massive graphs. In ACM-SIAM, editor, *12th Ann. Symp. on Discrete Algorithms (SODA)*, pages 912–921, 2001.

[27] T. Luczak. Sparse random graphs with a given degree sequence, in Random Graphs, vol. 2. A.M. Frieze, T. uczak eds. Wiley, New York, 1992. pages. 165-182.

[28] D. Magoni and J.-J. Pansiot. Analysis of the autonomous system network topology. *ACM SIGCOMM Computer Communication Review*, 31(3):26 – 37, July 2001.

[29] D. Magoni and J.-J. Pansiot. Internet topology modeler based on map sampling. In *Proceedings of ISCC'02, IEEE Symposium on Computers and Communications*, Italy, July 2002.

[30] D. Magoni and J.-J. Pansiot. Influence of network topology on protocol simulation. In *ICN'01 - 1st IEEE International Conference on Networking*, volume Lecture Notes in Computer Science, pages 762–770, July 9-13, 2001.

[31] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, pages 161–179, 1995.

[32] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. *Combin. Probab. Comput.*, pages 295–305, 1998.

[33] M.E.J. Newman, D.J. Watts, and S.H. Strogatz. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA*, 99 (Suppl. 1):2566–2572, 2002.

[34] J.-J. Pansiot and D. Grad. On routes and multicast trees in the internet. *ACM Computer Communication Review*, 28(1):41–50, 1998.

[35] R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, 2004.

[36] T. Petermann and P. De Los Rios. Exploration of scale-free networks. *To appear in Eur. Phys. J. B*, 2004.

[37] DIMES@home Project. http://www.cs.huji.ac.il/ eproj/available/dimes.

[38] Traceroute@Home project. University of Paris 6, coordinator: Timur friedman.

[39] P. De Los Rios. Exploration bias of complex networks. In *Proceedings of the 7th Conference on Statistical and Computational Physics Granada*, 2002.

[40] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP topologies with rocketfuel. In *Proceedings of ACM/SIGCOMM '02*, August 2002.

[41] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. On characterizing network hierarchy. Technical Report 03-782, Computer Science Department, University of Southern California, 2001. submitted.

[42] A. Vazquez, R. Pastor-Satorras, and A. Vespignani. Internet topology at the router and autonomous system level. [cond-mat/0206084].

[43] B.M. Waxman. Routing of multipoint connections. *IEEE Journal of Selected Areas in Communications*, pages 1617–1622, 1988.

[44] E.W. Zegura, K.L. Calvert, and M.J. Donahoo. A quantitative comparison of graph-based models for Internet topology. *IEEE/ACM Transactions on Networking*, 5(6):770–783, 1997.